基于文本内容特征选择的评论质量检测*

孟 园 王洪伟
(同济大学经济与管理学院 上海 210000)

摘要: 【目的】在有效提取多维特征基础上, 考察评论内容特征对评论质量检测的影响。【方法】基于评论文本的信息特征度量和情感倾向的混合性, 量化并抽取评论内容特征, 采用 GBDT 模型评估特征集合分类效果, 结合贪婪式特征选择算法识别有效内容特征, 分析其对评论质量检测的影响。【结果】将评论内容特征应用于评论质量检测任务中能取得较好的效果, 明显提升了实验准确率和召回率。【局限】实验对象主要是搜索型产品的评论数据, 未对其他享受型产品(如电影、音乐)等进行验证和比较。【结论】评论内容的信息增益、产品特征词的信息增益、评论客观情感倾向度、内容差异性对评论质量检测有明显作用。

关键词: 评论质量 信息特征 情感倾向 内容特征 贪婪式特征选择
分类号: G350

1 引 言

随着互联网技术的日益成熟, 消费者网络点评积极性逐渐增强, 网络上产生了数量庞大的评论数据。用户利用这些评论信息辅助购买决策的同时, 也饱受评论质量参差不齐、信息过载等问题的困扰, 仅依靠人工方法难以从海量的评论中识别出真正对用户有价值的信息, 迫切需要自动化方法辅助人们进行甄别, 因而对在线评论的质量进行检测具有重要的研究价值。

一些购物网站通过设置"有用性投票"对评论质量进行排序, 基于此, 学者普遍认为消费者对评论的感知有用性度量了评论的质量或效用, 有用性程度越高, 代表评论质量或效用越高, 因而评论质量、评论效用与评论有用性一般视为同等概念[1-2]。现有文献对评论质量的检测方法主要分为两种: 计量回归方法和监督学习方法。前者一般以元数据特征(如评论评分、评论者身份)或语言特征(如评论字数、词语数等)作为自变量, 评论有用性投票比例作为因变量, 考察哪些元数据特征或语言特征对评论质量影响显著。而后者则将评论质量的检测视为分类问题, 采取设置有用性投票比例阈值或人工标注方法生成有用评论训练模型, 利用最优模型自动识别高质量评论, 效果相对较好。由于评论质量受多种特征因素影响, 如何选择有效特征是评论质量检测的关键。目前, 国内研究对有用评论的特征选择集中在元数据特征、语言特征的等方面[2-6], 对文本内容特征的挖掘还不够深入, 较少涉及特征的贡献度和选择机制分析。

本文以梯度提升决策树模型(Gradient Boosting Decision Tree, GBDT)作为分类模型, 在提取多维特征基础上, 重点考察评论内容的信息特征和语义情感特征在分类模型上的表现, 进一步利用贪婪式特征选择算法识别有效的内容特征集合, 深入揭示多维评论特征的影响效果。

2 文献综述

2.1 影响评论质量的特征分析

现有研究中影响评论质量的特征大致可以分为三大类: 元数据特征、语言特征和评论内容特征。杨铭等[2]指出元数据特征与文本内容信息和文本语言特征无关, 评论评分、评论有用投票数、评论总投票数

通讯作者: 孟园, ORCID: 0000-0002-6595-8370, E-mail: nancymeng5544@163.com。

*本文系国家自然科学基金项目"中文语境下基于模糊本体的用户在线评论的情感分析"(项目编号: 70971099)和国家自然科学基金项目"在线评论对商家业绩的影响研究: 情感分析的视角"(项目编号: 71371144)的研究成果之一。

等是重要的元数据特征。Kim 等[7]研究表明评论发表距今的时间是显著影响评论质量的元数据特征。Ghose 等[8]认为，评论者相关信息是有效的元数据特征，例如评论者以往发表的评论数及有用率、评论者身份等。语言特征则主要是指从词频统计的角度发现评论的特征。如 Ghose 等[8]、Li 等[9]、Liu 等[10]指出主要的语言特征应包括评论字数、句子数、不同词性(名词、动词、形容词等)的词语数等。Chen 等[11]强调在评论所包含的名词中，产品属性名词的频次是重要的语言特征，高质量的评论中应包含一定数量的产品属性名词。从这些研究中，可以发现元数据特征和语言特征属于外在层面的评论特征，与此相对应的，内在层面的特征基于评论文本内容，消费者阅读评论后，能了解其他用户对产品的正面或负面的观点评价，从而对产品认知获取到一定程度的信息量，以消除对产品认知的不确定性。王伟等[12]指出正是这些从评论内容中获取的信息真正影响了消费者的购买意愿，聂卉等[1]重点验证了评论情感特征对评估评论效用具有较好效果，可见，评论内在特征是消费者判断评论质量的重要依据。

2.2 评论质量检测方法

已有研究主要采取计量方法和监督学习方法检测评论质量。计量方法研究一般以有用性投票比例作为评论质量的代理变量，比例越高，评论质量越高。如 Ghose 等[8]采用多元线性回归方法，对 DVD 产品的评论数据进行验证，得出评论者特征和评论语言特征对评论质量有显著正向影响。同样采用计量方法的还有文献[5-6]，分别得出评论字数、评论评分、评论长度等特征能影响评论质量。另一方面，监督方法将评论质量检测视为一个分类问题，通过人工标注或设置有用性投票比例阈值标注有用评论训练集，利用提取的特征集来测试和评估分类器效果，从而发现有效的评论特征，以自动识别高质量评论。如聂卉等[1]利用有用性投票比例作为评论质量代理指标，设置合理阈值生成有用性评论训练集，采用随机森林方法检测评论质量。另外，以人工标注获得训练集，Liu 等[10]采用支持向量回归、决策树等机器学习方法进行比较，以得到性能最优的分类模型。Chen 等[11]构造了多层支持向量机对评论质量进行分类。

2.3 研究述评与问题定义

由以上文献综述分析得出，特征选取对于评论质量的检测十分关键。现有研究大多关注语言特征和元数据特征等外在特征，虽有少数学者专门验证了文本内容情感特征的作用，但鲜有全面考察外在特征和内在特征对评论质量的影响。评论内容特征对评论质量的影响是否明显？特征选择顺序是否影响分类效果？这些问题仍然需要得到解答。针对现有研究的不足，本文的目标是采用 GBDT 监督学习方法，深入、全面挖掘影响评论质量的有效特征集合，考察评论内容特征对评论质量的影响。因而，研究过程从以下三个方面展开:

(1) 提取评论内容的内在特征，包括信息特征和语义情感特征。

(2) 采取 GBDT 分类方法和贪婪式特征选择算法，识别有效特征集合和最佳分类模型。

(3) 分类模型的性能评测和比较。

3 研究框架

本文将评论质量检测任务建模为二元分类问题，在对文本多维特征有效提取的基础上，采用梯度提升决策树模型(GBDT)和贪婪式特征选择算法进行最佳模型识别。对评论质量进行分类学习。GBDT 模型组合“基学习器(Base Learner)”，经多次迭代，每次迭代过程根据损失函数在梯度下降方向上建立决策树模型，使得相加的损失函数(Loss Function)最小，通过迭代改进能获得比基学习器更为良好的分类性能，在分类、回归等研究问题上表现优异[13-14]。本文研究的主要任务包括实验评论选取、分句、特征提取、特征选择、模型训练与模型识别、实验结果分析等过程，研究框架如图 1 所示。

3.1 文本内容特征提取

本文的重点是考察评论内容特征的效果，因此重点阐述评论内容相关特征的提取方法，着重从文本内容蕴含的信息特征和语义特征两方面，提取 8 个特征，如表 1 所示。

(1) 信息特征提取

①评论内容的信息量

从信息论的角度来看，评论 r 蕴含的信息量越大，这条评论对用户越有用。评论中不同的词语为评论有用性贡献不同的信息量，因而本文利用词语的信息增益量化评论 r 的信息量。

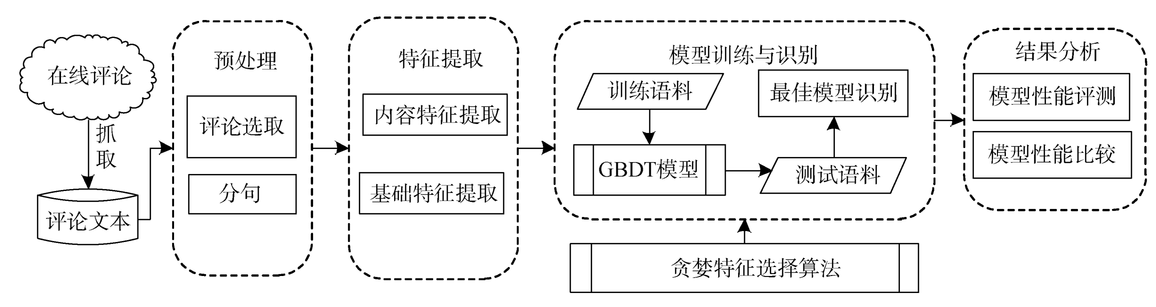


图 1　研究框架

表 1　各种内容特征概述

内容特征集合	特征定义	特征描述
信息特征(I)	I1: IGain(r)	评论 r 的信息量
	I2: IGain$_f$(r)	评论 r 包含的特征词信息量
	I3: Entropy(r)	评论 r 的信息熵
	I4: Perplexity(r)	评论 r 的困惑值
语义情感特征(S)	S1: ObjDegree(r)	评论 r 的客观情感倾向度
	S2: DevObj(r)	评论 r 的客观情感倾向度偏差
	S3: PosDegree(r)	评论 r 的正向情感倾向度
	S4: DevPos(r)	评论 r 的正向情感倾向度偏差

对于给定的在线评论集合 R，以 C=(c_1, c_2)表示评论空间有用性类别，其中 c_1 表示有用类别，c_2 表示无用类别，则判断二分类系统所需的信息熵总量 H(C)为：

$$H(C)=-\sum_{i=1}^{2}P_r(c_i)\log P_r(c_i) \quad (1)$$

其中，$P_r(c_i)$为系统中类别 c_i 的出现概率。

评论 r 由多个不同的词组成，考虑评论中的某个词语 t，其可能的取值为两种，出现或不出现，分别用 w 和 $\overline{w}$ 表示，则当 t 出现的条件下(即 t 取值为 w)，系统包含的信息熵 H(C | w) 为：

$$H(C\mid w)=-\sum_{i=1}^{2}P_r(c_i\mid w)\log P_r(c_i\mid w) \quad (2)$$

其中，$P_r(c_i\mid w)$ 为 t 出现的评论中，类别 c_i 出现的概率。

同理，可以得到 t 不出现(即 t 取值为 $\overline{w}$)的条件下，系统包含的信息熵 $H(C\mid\overline{w})$。考虑 t 两种不同取值条件下为系统带来的信息增量，即 t 的信息增益 G(t)为：

$$G(t)=H(C)-P_r(w)H(C\mid w)-P_r(\overline{w})H(C\mid\overline{w}) \quad (3)$$

其中，$P_r(w)$ 表示 t 的出现概率，$P_r(\overline{w})$ 表示 t 的不出现概率。

由于 G(t)考虑的是词 t 在两个类别(c_1 和 c_2)整体上的贡献度之和，考虑到有用评论能帮助用户消除对产品不确定性的认知，而无用评论不仅无法给用户购买决策提供支持，可能还会影响用户对产品的正确判断，因而词 t 在有用类别不同评论中，其信息增益方向是不同的。为了更好地体现 t 在两个类别中信息增益的差异，借鉴文献[15]对词 t 的信息增益进行改进，在 t 出现的所有评论中，比较有用类别和无用类别的出现概率，即 $P_r(c_1\mid w)$ 和 $P_r(c_2\mid w)$，如果前者大于后者，则词语 t 代表正向信息增益，反之则代表负向信息增益。改进后的词 t 的信息增益 IG(t) 表示为：

$$IG(t)=\begin{cases}G(t) & 若 P_r(c_1\mid w)>Pr(c_2\mid w)\\ -G(t) & \text{otherwise}\end{cases} \quad (4)$$

由此，评论 r 的信息量 IGain(r)表示为 r 中所有词语的信息增益之和，公式如下所示：

$$IGain(r)=\sum_{t\in r}IG(t) \quad (5)$$

文献[12]表明评论中产品特征词相比其他词语，其信息增益对于用户判断评论质量作用更大，为此，考察每个产品特征词 f 提供的信息增益 IG(f)，并提取每条评论中所有特征词的贡献的信息量 $IGain_f$，其计算公式如下：

$$IGain_f(r)=\sum_{f\in r}IG(f) \quad (6)$$

②评论内容的差异性

文献[16]指出，内容越相似的评论越有可能是虚假评论，这反映出评论内容与其他评论内容的差异性影响用户对评论质量的感知。贾里尼克在文本信息熵基础上，定义了困惑值的概念，两者同时使用，可以度量一条评论与其他评论在内容上的差异性[17]。对于评论集合 R，如果评论 r 和其他评论内容差异越大，则评论 r 的信息熵和困惑值就越大。假设评论 r 由一连串特定顺序排列的词 $w_1,w_2,\cdots w_n$ 组成，$p(w_i)$ 为 r 中词语 w_i 出现的概率，则该条评论的信息熵 Entropy(r)和困惑值 Perplexity(r)的表示如下：

$$Entropy(r)=-\sum_{w_i\in r}p(w_i)\log p(w_i) \quad (7)$$

$$Perplexity(r)=2^{Entropy(r_i)} \quad (8)$$

本文以每个产品型号对应的评论子集分别作为训练语料，构建 unigram 统计语言模型，再使用训练模型计算对应子集内每条评论的信息熵和困惑值。

(2) 语义情感特征提取

评论中经常呈现出混合观点形式，既包含正面或负面情感，也有主观或客观情感。通常，评论观点的正负面情感倾向由评论中的观点词极性来决定，而评论的主客观情感倾向则由评论者对商品属性点评与商家描述的一致性程度度量[18]，即评论文本与商家描述内容越相似，说明评论的用语比较正式，评论文本趋向

于客观。例如评论句:

"这款产品性能挺优越的，外观上也非常小巧漂亮。性价比一般吧，因为价格有点高。总体来说，我还是非常喜欢的!"

从情感极性上来看，这条评论表达了正向和负向两种情感倾向，但整体而言情感表达是正向的，而从内容的主客观性上分析，前两句评论相比后一句评论，则更接近于客观的评论。为全面度量评论中情感的混合性对评论质量的影响，以下定义客观情感倾向度 ObjDegree 及其偏差 DevObj、正向情感倾向度 PosDegree 及其偏差 DevPos 等 4 个特征项。

①客观情感倾向度及其偏差

以评论子句为单位，考察评论内容与产品描述文本的余弦相似性，判断其客观性。利用文本词语的 tf-idf 权值对评论子句 s 和商品描述 d 分别进行向量表示，计算两者的余弦相似度 sim(s,d)，设定阈值 λ 判断评论子句的客观性。以 s^{+}表示 r 中客观的评论子句，total(r)为评论 r 中的评论子句总数，则评论 r 客观情感倾向度计算公式为:

$$\mathrm{ObjDegree}(r)=\frac{\mathrm{count}(s^{+})}{\mathrm{total}(r)} \tag{9}$$

对于同一产品 p 的所有评论平均客观情感倾向度，均匀地反映整体主客观观点句比例的稳定值，将评论 r 的客观情感倾向度与整体均值进行比较，偏差越大，说明评论 r 中越有可能呈现一致性观点(都是客观或都是主观观点)，偏差越小，说明评论 r 越有可能呈现主客观混合观点(既有客观观点也有主观观点)。

因此，基于产品 p 所有评论的平均客观情感倾向度，定义评论 r 的客观情感倾向偏差，表示评论主客观情感混合程度，其计算公式为:

$$\mathrm{DevObj}(r)=|\mathrm{ObjDegree}(r)-\mathrm{Avg}(\sum_{r_i\in R}\mathrm{ObjDegree}(r))| \tag{10}$$

②正向情感倾向度及其偏差

评论一般由多个观点子句构成，以正向观点子句的占比代表评论的正向情感倾向度，占比越大，说明整条评论偏向于正向情感，反之则偏向于负向情感，因而，正向情感倾向度表达了评论的情感极性特征。以评论 r 中的子句为单位，判断其情感极性。本文采取机器学习方法对评论子句进行情感极性分类。从实验语料中选取 5 星评分和 1 星评分评论各 1 000 条构建情感分类器。根据卡方统计值选择前 1 500 个单词(unigram)和双词(bigram)作为文本情感极性分类特征项[1]，在 Python 环境下选择分类效果最好的 BernoulliNB 作为分类器对评论子句正负情感极性进行判别。以 r^{+}表示 r 中正向的评论子句，total(r)为评论 r 中的子句总数，则评论 r 的正向情感倾向度计算如下:

$$\mathrm{PosDegree}(r)=\frac{\mathrm{count}(r^{+})}{\mathrm{total}(r)} \tag{11}$$

同理，正向情感倾向偏差度量了评论 r 的正负情感混合程度特征，其计算如下所示:

$$\mathrm{DevPos}(r)=|\mathrm{PosDegree}(r)-\mathrm{Avg}(\sum_{r\in R}\mathrm{PosDegree}(r))| \tag{12}$$

3.2　特征选择

(1) 基础特征模板

由于评论质量与评论元数据特征(Meta)和语言特征(Lan)密切相关，为此将元数据特征和语言特征作为基本特征集合，构建分类模型的特征模板。根据文献[3,5,7-8,10-11]中研究结论，提取 6 个有效元数据特征(M1-M6)和 3 个语言特征(L1-L3)，如表 2 所示:

表 2　基础特征集合

特征	描述
M1	评论 r 的有用投票率，评论 r 获得的有用投票数除以总投票数
M2	评论 r 获得的有用投票数
M3	评论 r 对应的用户评分
M4	评论 r 发表至今的时间
M5	评论 r 对应的评论者排名
M6	评论 r 对应的评论者以往评论的平均有用率
L1	评论 r 包含的字数
L2	评论 r 包含的词语数
L3	评论 r 的产品特征词数

(2) 贪婪式特征选择算法

为了能够从提取的文本内容相关特征集中分别选择有利于评论质量检测的特征集，以元数据特征和语言学特征为基本特征集合，以提取的内容特征为候选特征集合，采用贪婪式特征选择算法[19]和 GBDT 分类模型进行特征选择。主要思路为：根据每个候选内容特征在开发集 DevData 上对分类任务的贡献度大小，每次选取贡献度最大的特征加入基本特征集合，当从剩余候选特征集中添加任意特征时，导致开发集的分类评价指标下降或剩余候选特征集为空时，算法终止。算法的执行流程如下:

输入：读入所有特征集合 F_{all}={M1~M6,L1~L3,I1~I4,S1~S4}

输出：有效特征集合 F_{select}={set of selected features}，M_{select}={selected model}

1: 初始化基础特征集合、候选内容特征集合，F_{select}={ M1~M6, L1~L3},$F_{can}=F_{all}-F_{select}$

2: 训练模型，得到初步分类性能 M_{select}=GBDT_Train(F_{select}), E_{select}=Evaluate(M_{select}, DevData)

3: 对文本内容特征进行选择

4: **loop**

```
5:      for each feature f_i in F_can do
6:          F_i= F_select U f_i
7:        M_i= GBDT_Train(F_i)
8:        E_i= Evaluate(M_i, DevData)
9:    end for
10:       E_max=Max(E_i)
11:       If E_max> E_select then
12:           F_select= F_select U f_max
13:           M_select=M_max
14:       E_select= E_max
15:           F_can= F_can-f_max
16:    end if
17:       if F_can= = ∅ or E_max<= E_select then
18:             return F_select, M_select
19:       end if
20: end loop
```

其中, 算法第 5-9 行为每次从候选特征集合 F_{can} 中选择一个特征 f_i 加入有效特征集合 F_{select}, 执行分类模型并记录其对应分类指标 E_i; 算法第 10-16 行为比较当前每个特征 f_i 对应的分类指标, 确定最大贡献度的 f_i 和其加入有效特征集合的顺序。

4 实验设计

4.1 实验数据及标注标准

利用爬虫程序抓取中文亚马逊网站的数码相机的相关评论信息和产品信息, 采集评论文本、评论元数据信息和产品描述文本, 数据采集截止时间为 2013 年 9 月 2 日, 评论发表时间跨度为 2009 年 1 月 7 日到 2013 年 9 月 1 日, 共采集了 15 327 条评论。选择其中评论总数大于 50 条的产品作为实验对象, 去除重复、广告评论等预处理操作后, 得到 10 568 条有效评论数据, 涵盖 10 个相机品牌、67 个型号的产品。具体统计信息如表 3 所示:

表 3 评论数据特征统计

评论相关属性	最小值	最大值	平均值
评论字数	1	3 296	62.93
评论词语数	1	1 961	38.13
评论产品特征词数	0	192	3.33
评论有用性投票数	0	1 122	2.01
评论有用率	0	1	0.20
评论评分	1	5	4.38
评论发表时间(log)	0.47	3.23	2.51
评论者排名(log)	1.77	6.64	5.90
评论 r 对应的评论者以往评论的平均有用率	0	1	0.38

鉴于有用性投票率有较大偏差, 借鉴文献[20], 对评论质量进行人工标注。邀请两名数码产品资深用户对实验数据进行独立标注。标注者逐条阅读所有评论, 并回答问题"该评论内容对您了解产品或购买产品有用吗?"。除了评论文本外, 还提供标注者评论对应的产品简要描述。经过对标注结果的 Cohen-Kappa 检验, 两名标注者的标注结果 Kappa 值达 83.45%, 可见标注者对于实验数据的质量评价标准达到了较高的一致性。以标注者 1 的标注结果训练和测试模型。最终获得 5 307 条高质量评论和 5 261 条低质量评论。

4.2 模型及评价指标

采用GBDT模型对评论质量进行分类, 经过测试, 分类中建立 25 棵树模型能达到最优分类效果。将实验数据按 4:1 分成训练集和测试集进行特征选择, 采用平均准确率、平均召回率和平均 F1 值作为评价标准, 分别记为 P、R、F1。

4.3 特征抽取结果

实验机器是 Win7 32bit 操作系统, 内存 4GB。使用 Python 语言编写程序, 在 Python2.7.3 下完成所有程序编写及测试。以下为部分文本内容特征抽取结果。

(1) 词语信息增益

利用实验数据的评论质量标注结果, 对评论进行分词、去停用词操作后, 计算评论集合词语(仅计算词性为 n、v、a、d、vn)的信息增益。有效词语共计 11 729 个, 部分词语的信息增益计算结果如表 4 所示:

表 4 部分词语信息增益计算结果

WordID	词语	正向信息增益	WordID	词语	负向信息增益
1	镜头	0.026199	11	签单	–0.000201
2	电池	0.019720	12	时机	–0.000204
3	拍	0.019718	13	拍下	–0.000207
4	机身	0.018958	14	正品	–0.000212
5	快门	0.017846	15	涨价	–0.000270
6	照片	0.017434	16	骗	–0.000274
7	清晰	0.016422	17	看上	–0.000286
8	功能	0.016267	18	不贵	–0.000349
9	屏幕	0.015868	19	自毁	–0.000381
10	对焦	0.015324	20	帮别人	–0.000467

(2) 部分特征抽取结果

对实验语料进行分词、分句等操作后, 按 3.1 节所述特征提取方法提取特征, 部分评论内容特征提取结果如表 5 所示。

表 5　部分内容特征提取结果

ReviewID	IGain	$IGain_f$	Entropy	Perplexity	ObjDegree	DevObj	PosDegree	DevPos
1	0.024422	0.002269	8.849878	461.401121	0.400000	0.020892	0.200000	0.355263
2	0.031540	0.008544	8.301512	315.503415	0.222222	0.156886	0.000000	0.555263
3	0.093255	0.036050	8.495248	360.848181	0.466667	0.087559	0.066667	0.488597
4	0.026955	0.008158	8.554623	376.008819	0.800000	0.420892	0.200000	0.355263
5	0.199504	0.022060	8.115530	277.343496	0.350000	0.029108	0.400000	0.155263
6	0.043566	0.002513	8.069165	268.572038	0.375000	0.004108	0.625000	0.069737
7	0.054861	0.000000	9.680146	820.378626	0.500000	0.120892	0.000000	0.555263
8	0.014244	0.000000	8.913717	482.276570	0.200000	0.179108	0.000000	0.555263
9	0.137508	0.076782	7.904827	239.656882	0.666667	0.287559	0.750000	0.194737
10	0.017129	0.000000	8.130206	280.179206	0.250000	0.129108	0.500000	0.055263

5　分类实验结果

(1) 文本内容特征的效果

以元数据特征和语言特征作为基础特征，然后依次加入文本内容特征。表 6 显示了分别加入单个文本内容特征的效果。

表 6　加入单个文本内容特征的效果

特征	P(%)	R(%)	F1(%)
Meta+Lan	72.92	74.20	73.56
+I1	77.15	74.56	75.83
+I2	76.56	75.88	76.22
+I3	76.60	76.06	76.33
+I4	76.60	76.06	76.33
+S1	76.60	76.06	76.33
+S2	76.51	75.97	76.24
+S3	76.53	76.06	76.30
+S4	76.53	76.06	76.30

从表 6 可以看出，加入单个内容特征后，评论质量的分类准确率和召回率都有一定程度的提高，F1 可以提高近 3 个百分点，验证了内容特征对评论质量检测的有效性。但依次加入内容特征后，分类指标值呈现先上升后下降的趋势，说明有些特征项的效果不明显，因此有必要进行特征选择，去除没有帮助的特征项。根据 3.2 节的特征选择算法得出，依次按照{I1,I2,S1,I3,I4}特征组合顺序，可以达到最好的分类效果。表 7 显示了利用贪婪式特征选择算法所选择的特征组合的分类效果，因此，将特征组合{I1,I2,S1,I3,I4}及其顺序作为最终有效的评论内容特征集合。

表 7　贪婪式特征选择结果

特征	P(%)	R(%)	F1(%)
Meta+Lan	72.92	74.20	73.56
+I1	77.15	74.56	75.83
+I2	76.56	75.88	76.22
+S1	76.60	76.06	76.33
+I3	76.67	76.06	76.36
+I4	76.67	76.06	76.36

从表 7 可以看出，过滤了冗余后的特征项中，仅有客观情感倾向度(S1)为有效情感特征，说明正式、客观的评论内容能影响评论质量。而所有信息特征对评论质量都具有明显作用，其中作用最大的是整体评论的信息量(I1)以及产品特征词的蕴含的信息量(I2)，其次分别是度量评论内容差异性的信息熵(I3)和困惑值(I4)。整条评论提供的信息量(I1)能帮助用户了解产品信息，产品特征词给用户判别评论质量提供了更有价值的信息，从而利于判别评论质量。评论内容的差异性对于评估评论质量起着非常关键的作用，这也间接验证了文献[16]的结论，即越相似的评论，越有可能是垃圾评论的论断。

(2) 基于有效特征集的模型比较

为考察 GBDT 模型的分类表现，基于有效特征集，与 Ghose 等[8]采用的随机森林模型(Random Forest, RF)进行比较。此外，与基本决策树模型(Decision Tree, DT)进行比较，考察梯度提升优化效果，实验比较结果如图 2 所示。可以看出，GBDT 模型方法与 RF 模型、DT 模型相比，准确性和召回率都有显著提高。整体来看，相比 DT 模型，F1 可以提高约 9 个百分点，说明

GBDT 优化效果较好；同时，相比 RF 模型，F1 也提高了约 2.3 个百分点，说明模型性能表现良好。

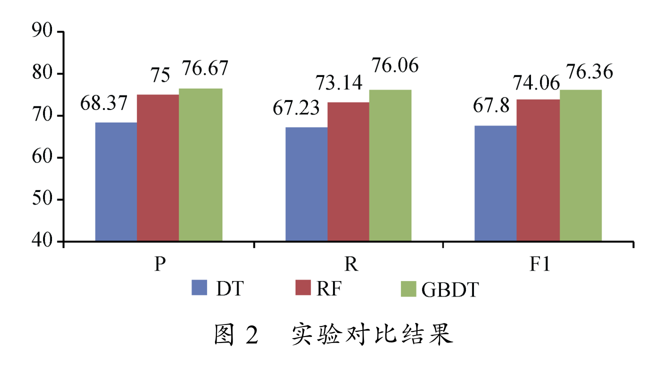


图 2　实验对比结果

6　结　语

本文主要介绍了文本信息特征和语义情感特征在评论质量检测中的应用效果，研究结果表明，经过贪婪式特征选择算法按一定顺序选择特征项后，GBDT 能在经过选择特征集上取得最佳分类性能，其分类效果优于决策树模型和随机森林模型，验证了特征提取和特征选择的有效性，从而更有效地帮助商家自动识别高质量评论。

未来将继续搜索其他有效的内容特征，进一步提高和完善文本特征在评论质量监测中的应用。

参考文献：

[1] 聂卉，容哲．面向评论效用评估的文本情感特征提取[J]．现代图书情报技术，2015 (7-8): 104-112. (Nie Hui, Rong Zhe. Review Helpfulness Prediction Research Based on Review Sentiment Feature Sets [J]. New Technology of Library and Information Service, 2015(7-8): 104-112.)

[2] 杨铭，祁巍，闫相斌，等．在线商品评论的效用分析研究[J]．管理科学学报，2012,15(5): 65-75. (Yang Ming, Qi Wei, Yan Xiangbin, et al. Utility Analysis for Online Product Review [J]. Journal of Management Science in China, 2012, 15(5): 65-75.)

[3] 高雅，李红，施慧斌．在线评论投票数的影响因素研究[J]．中国管理信息化，2012，15(17): 88-91. (Gao Ya, Li Hong, Shi Huibin. The Research of the Impact Factors of the Online Review Votes [J]. China Management Informationization, 2012, 15(17): 88-91.)

[4] 严建援，张丽，张蕾．电子商务中在线评论内容对评论有用性影响的实证研究[J]．情报科学，2012，30(5): 713-719. (Yan Jianyuan, Zhang Li, Zhang Lei. An Empirical Study of the Impact of Review Content on Online Reviews Helpfulness in E-Commerce [J]. Information Science, 2012, 30(5): 713-719.)

[5] 杨爽．信息质量和社区地位对用户创造产品评论的感知有用性影响机制——基于 Tobit 模型回归[J]．管理评论，2013，25(5): 136-143,154. (Yang Shuang. The Impact Mechanism of Information Quality and Community Status on Perceived Usefulness for User-Generated Product Reviews——Tobit Regression Analysis [J]. Management Review, 2013, 25(5): 136-143, 154.)

[6] 殷国鹏．消费者认为怎样的在线评论更有用?——社会性因素的影响效应[J]．管理世界，2012(12): 115-124. (Yin Guopeng. What is the Kind of Online Reviews that Consumer Think are More Useful? The Effect of Social Factors Influence [J]. Management World, 2012(12): 115-124.)

[7] Kim S M, Pantel P, Chklovski T, et al. Automatically Assessing Review Helpfulness [C]. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia. Stroudsburg, PA, USA: ACL, 2006: 423-430.

[8] Ghose A, Ipeirotis P G. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(10): 1498 -1512.

[9] Li F, Zhang Y L, Dang Y, et al. Analyzing Sentiments in Web2.0 Social Medial Data in Chinese: Experiments on Business and Marketing Related Chinese Web Forums [J]. Information Technology Management, 2013(14): 231-242.

[10] Liu Y, Jin J, Ji P, et al. Identifying Helpful Online Reviews: A Product Designer's Perspective [J]. Computer-Aided Design, 2013, 45(2): 180-194.

[11] Chen C C, Tseng Y-D. Quality Evaluation of Product Reviews Using an Information Quality Framework [J]. Decision Support Systems, 2011, 50(4): 755-768.

[12] 王伟，王洪伟．特征观点对购买意愿的影响：在线评论的情感分析方法[J]．系统工程理论与实践，2016，36(1): 63-76. (Wang Wei, Wang Hongwei. The Influence of Aspect-based Opinions on User's Purchase Intention Using Sentiment Analysis of Online Reviews [J]. Systems Engineering——Theory & Practice, 2016, 36(1): 63-76.)

[13] Ayaru L, Ypsilantis P-P, Nanapragasam A, et al. Prediction of Outcome in Acute Lower Gastrointestinal Bleeding Using Gradient Boosting [J]. PLoS ONE, 2015, 10(7). DOI: 10.1371/journal.pone.0132485.

[14] Semanjski I, Gautama S. Smart City Mobility Application-Gradient Boosting Trees for Mobility Prediction and Analysis Based on Crowd Sourced Data [J]. Sensors, 2015, 15(7): 15974-15987.

[15] Zhang R, Tran T. An Information Gain-based Approach for Recommending Useful Product Reviews [J]. Knowledge and Information Systems, 2011, 26(3): 419-434.

[16] Jindal N, Liu B. Review Spam Detection [C]. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada. New York, NY, USA: ACM, 2007: 1189-1190.

[17] 吴军. 数学之美[M]. 北京: 人民邮电出版社, 2012:60-64. (Wu Jun. The Beauty of Mathmatics [M]. Beijing: Posts and Telecom Press, 2012: 60-64.)

[18] Pang B, Lee L. A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts [C]. In: Proceeding of the 42nd Annual Meeting of the Association for Computational Linguistic (ACL). Morristown, NJ, USA: ACL, 2004: 271-278.

[19] Jiang Z P, Ng H T. Semantic Role Labeling of NomBank: A Maximum Entropy Approach [C]. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: ACL, 2006: 138-145.

[20] Liu J, Cao Y, Lin C-Y, et al. Low-Quality Product Review Detection in Opinion Summarization [C]. In: Proceedings of the 2007 Joint Conference on EMNLP-CoNLL. ACL Press, 2007: 334-342.

作者贡献声明:

孟园: 采集、清洗和分析数据, 论文起草及最终版本修订;
王洪伟: 论题拟定, 提出研究思路, 设计研究方案, 修改完善论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 孟园, 王洪伟. quality_classify.py. 评论质量贪婪式特征选择和模型比较算法.

[2] 孟园, 王洪伟. exp_data_initial.xlsx. 亚马逊原始抓取数据.

[3] 孟园, 王洪伟. classifier.pkl. 文本情感极性分类器.

[4] 孟园, 王洪伟. segged_pos_words.txt. 评论分词数据.

[5] 孟园, 王洪伟. sub_sentence.txt. 评论分句数据.

[6] 孟园, 王洪伟. Refine_Igain_word.txt. 改进评论词语信息增益数据.

[7] 孟园, 王洪伟. all_feature_value.txt. 评论特征提取数据.

收稿日期: 2015-12-09
收修改稿日期: 2016-02-07

Evaluating Online Reviews Based on Text Content Features

Meng Yuan Wang Hongwei
(School of Economics and Management, Tongji University, Shanghai 210000, China)

Abstract: [**Objective**] This paper aims to effectively extract multi-dimensional characteristics of online reviews and then examine the impact of text content to the review quality evaluation. [**Methods**] First, we quantified and extracted content features based on the textual and sentimental message from the reviews. Then, adopted the GBDT model to evaluate the influence of feature sets to classification results, along with greedy feature selection procedure to identify the most effective content features. Finally, we examined the influences of these features. [**Results**] The proposed method could improve the performance of review quality evaluation tasks, especially the recall and precision of the new system. [**Limitations**] Our research focused on review data from search services, and did not investigate products like movies and music. [**Conclusions**] The information gained from reviews and product feature words, degree of sentimental objectiveness, and differences among review contents all posed important effects to review quality evaluation.

Keywords: Review quality Information feature Sentiment orientation Review content Greedy feature selection